Glòria Vàzquez
Universitat de Lleida

Víctor Siurana 1
25003-Lleida
gvazquez@dal.UdL.es

Laura Alonso
Universidad Nacional de Córdoba

FaMAF - UNC
Haya de la Torre s/n
Córdoba - Argentina
alemany@famaf.unc.edu.ar

Joan A. Capilla
Universitat de Lleida

Victor Siurana 1
25003-Lleida
jcapilla@dal.udl.es

Irene Castellón
Universitat de Barcelona

Gran Via de les Corts Catalanes 585
08007-Barcelona
icastellon@ub.edu

Ana Fernández
Universitat Autònoma de Barcelona

Emprius 2
08202- Sabadell
ana.fernandez@uab.es

Resumen: En este artículo presentamos el desarrollo del proyecto SenSem (BFF2003-06456), que tiene como objetivo describir y representar el comportamiento léxico, sintáctico y semántico de los verbos del español. En el desarrollo de este proyecto se están construyendo dos recursos: un corpus de oraciones asociadas a su interpretación sintáctico-semántica y un léxico donde cada sentido verbal se asocia a un conjunto de ejemplos anotados del corpus.

Abstract: This paper presents the development of the SenSem project (BFF2003-06456),which aims at describing and representing the lexical, semantic, and syntactical behaviour of Spanish verbs. Two resources are being developed in the course of this project: a corpus of sentences associated to their syntactico-semantic interpretation, and a lexicon where each verb meaning is linked to a number of annotated examples from the corpus.

Palabras clave: anotación de corpus, lexicón verbal, semántica oracional, sintaxis y semántica
Keywords: corpus annotation, verb lexicon, sentence semantics, syntax and semantics

1 Introducción

El objetivo del proyecto SenSem es la construcción de un banco de datos de verbos del español. Dicho banco de datos se compone de un léxico donde cada sentido verbal está asociado a un conjunto de ejemplos del corpus analizados y anotados a diferentes niveles. Este banco de datos reflejará el comportamiento sintáctico-semántico de 250 verbos del español.

Con este objetivo, se han seleccionado 250 verbos frecuentes del español, y se ha constituido un corpus de 25.000 oraciones con 100 ocurrencias de cada verbo. El corpus contiene 750.000 palabras, de las cuales 350.000 están bajo el alcance sintáctico de alguno de los verbos que han sido anotados.

La anotación se lleva a cabo en tres niveles: el verbo como unidad léxica, los constituyentes de la oración y la oración como un todo. Aunque somos conscientes de que se requeriría un mayor número de frases anotadas para extraer datos fiables a nivel estadístico, creemos que este corpus constituye un buen punto de partida para el estudio de los fenómenos que en él se describen.

El proceso de anotación, que es manual, incluye aspectos sintácticos y semánticos. Respecto a la sintaxis, describimos las categorías sintácticas no léxicas, los núcleos de dichas estructuras y las funciones sintácticas. Consideramos que la aportación más relevante del proyecto se centra en la anotación semántica. Nuestro interés va más allá de la pieza léxica, por lo que, además de desambiguar los sentidos verbales, se ha llevado a cabo también la interpretación de papeles semánticos y la anotación de varios tipos de semántica oracional. Este último aspecto es precisamente lo que caracteriza nuestro proyecto respecto de algunos de los que se están desarrollando actualmente en español, que, como en SenSem, trabajan la semántica más allá del ámbito léxico (Subirats and Petruck, 2003; García De Miguel y Comesaña, 2004, Palomar et al. 2004).

A partir del análisis de un número significativo de ejemplos, se está sistematizando y codificando en un léxico el comportamiento prototípico de los sentidos del verbo. Esta descripción verbal asociada al sentido, se centrará en las características de la interfaz sintáctico-semántica, e incluirá datos relativos tanto al propio verbo como a las estructuras en que participa.

La conjunción de toda esta información proporcionará una descripción muy detallada de la interfaz sintáctico-semántica al nivel de la oración, útil para las aplicaciones que requieren una comprensión de oraciones más allá del análisis estrictamente sintáctico. En los campos de la comprensión automática, de la representación semántica y de los sistemas automáticos de aprendizaje, un recurso de este tipo es especialmente valioso, dado que como resultado se obtendrán las diferentes estructuras sintácticas asociadas a información semántica de diferentes niveles: léxico, sintagmático y oracional.

En el resto del artículo describiremos más detalladamente el proceso de anotación y presentaremos ejemplos de este proceso (sección 2). En las secciones 3 y 4 se presentan la metodología y las herramientas que se han utilizado, respectivamente. Por último, finalizamos con algunas conclusiones obtenidas hasta el momento.

2 Niveles de anotación

En el análisis de las oraciones tenemos en cuenta únicamente los constituyentes directamente relacionados con el predicado verbal. Los elementos que están más allá del alcance del verbo quedan excluidos en el proceso de anotación. En (1) podemos ver un ejemplo del alcance de la anotación.

(1) Con estas frases el Papa respondió de manera indirecta, a quienes **afirman** que debido a sus enfermedades debería renunciar al Papado.

Como podemos observar en este ejemplo del verbo *afirmar*, la anotación que estamos realizando ignora los participantes de la oración principal y sólo se tienen en cuenta los elementos de la frase subordinada, ya que son los que están al alcance del verbo que estamos anotando.

En el caso de que se anotara el verbo *responder*, se anotaría la frase subordinada como un único constituyente sin analizar la relativa internamente, ya que en cada oración se trabaja sobre una única unidad verbal.

Las oraciones se anotan a tres niveles: el léxico, en el que se da cuenta del sentido verbal, el de constituyentes, en el que se caracterizan los participantes de la oración, y el oracional, en el que se caracteriza el significado del conjunto. Veamos a continuación cada uno de estos niveles.

2.1 Nivel léxico

En el nivel léxico, cada ejemplo se asigna a un sentido verbal. Para ello hemos desarrollado un léxico verbal en el que se han distinguido *a priori* los posibles sentidos para cada verbo. Este léxico se ha desarrollado a partir del trabajo anterior de miembros del grupo en la creación de léxicos multilingües (Fernández et al. 2002). Las fuentes lexicográficas que se han utilizado para establecer el inventario de sentidos de cada verbo han sido principalmente el Diccionario de la Real Academia de la Lengua Española y el Diccionario Salamanca de la Lengua Española. En la construcción de este léxico, no se tuvieron en cuenta los significados menos frecuentes y los de usos arcaicos o muy restringidos.

Además, en este léxico inicial se describe la clase eventual del predicado, distinguiendo entre eventos, procesos y estados, y los roles semánticos básicos que caracterizan a los participantes verbales.

Por otro lado, se asocia a cada unidad una lista de sentidos verbales sinónimos y los synsets relacionados en EuroWordNet (Vossen 1999). Por lo que respecta a la conexión de nuestro léxico con WordNet, cabe señalar que dado que nuestro recurso tiene un enfoque más lingüístico, normalmente el enlace es de un sentido de nuestro corpus a varios de los recogidos en WordNet.

2.2 Nivel de constituyentes

En este nivel, que se sitúa claramente en la interfaz entre la sintaxis y la semántica, se anota, para cada participante de la frase, la categoría sintagmática. (SN, SP, etc.), la función sintáctica (sujeto, objeto directo, objeto indirecto, objeto preposicional, etc.) y el tipo de relación argumental con el verbo (argumento o adjunto).

Los argumentos se definen como participantes del predicado y, desde nuestra perspectiva, forman parte de la semántica léxica del verbo, por lo que también se les asocia un rol semántico (agente, tema, iniciador, etc.).

También se anotan los núcleos de los argumentos, ya que esta información nos es útil para adquirir las preferencias selectivas de cada sentido verbal.

Además, se ha incluido información relevante sobre unidades que pueden alterar alguna interpretación o bien que consideramos interesantes para trabajos futuros. Un ejemplo de este tipo de información es la polaridad negativa.

2.3 Nivel de semántica oracional

En este nivel agrupamos diversos aspectos que caracterizan el significado oracional, mediante atributos que expresan la semántica oracional como anticausativa, antiagentiva, impersonal, reflexiva o recíproca.

Este tipo de información es útil para especificar la estructura argumental de cada unidad verbal, ya que puede suceder que dos configuraciones que hayan sido asociadas al mismo sentido y hayan sido anotadas con estructuras sintácticas equivalentes, tengan un significado oracional diferente. Según esto, la relación entre roles semánticos y funciones sintácticas será distinta, y es muy probable que en otra lengua se expresen con distintas formas sintácticas.

3 Metodología

3.1 Composición del corpus

El corpus SenSem se compone de 25.000 oraciones del español anotadas a nivel sintáctico-semántico. Estos 25.000 ejemplos ilustran el comportamiento de los 250 verbos usados con más frecuencia en el español, según datos estadísticos extraídos de un corpus periodístico de más de 13 millones de palabras. Se han extraído de forma aleatoria 100 ejemplos correspondientes a cada verbo de un corpus de la versión electrónica de *El Periódico de Catalunya*. Los ejemplos excluyen los usos perifrásticos de los verbos, así como expresiones idiomáticas y colocaciones.

En total, el corpus contiene 750,000 palabras, de las cuales 350,000 están dentro del alcance de algún verbo que ha sido anotado. En este momento se ha finalizado la anotación del corpus y se está llevando a cabo el proceso de revisión, por lo que estas cifras pueden variar ligeramente.

Sentido	Definición
Alcanzar 1	Poder tocar algo o a alguien
Asegurar 1	Fijar algo dándole firmeza
Casar 1	Celebrar la cerimonia civil o eclesiástica uniendo a dos personas
Dejar 1	Descuidar algo en algún lugar
Echar 1	Tirar algo, a veces impulsándolo, para que llegue a un sitio.
Reclamar 1	Mostrar el desacuerdo ante algo y hacer las gestiones necesarias para que eso cambie

Tabla 1: Algunos sentidos sin representación en el corpus

Núm. ejemplos	% de sentidos	Núm. sentidos
Más de 90	6,2%	71
Entre 90 y 75	4,3%	49
Entre 75 y 50	6,1%	69
Entre 50 y 25	9,0%	103
Entre 25 y 10	11,8%	135
Entre 10 y 5	7,1%	81
Entre 5 y 2	13,5%	154
1	7,5%	86
0	34,0%	392

Tabla 2: Distribución de los ejemplos entre sentidos

La validez de la frecuencia de estos verbos es relativa, ya que se han extraído de un corpus periodístico, que, aunque se caracteriza por incluir textos escritos en lenguaje estándar y de diferentes temáticas, no es un corpus de referencia. Una consecuencia de ello es que no se han encontrado representados en el corpus todos los sentidos definidos para cada verbo, algunos de los cuales parecen, a simple vista, muy habituales (v. tabla 1). En la tabla 2 presentamos los porcentajes sobre el volumen de ejemplos y los sentidos. Así pues, tenemos previsto consultar otras fuentes para poder ampliar el lemario con los verbos más frecuentes extraídos a través de un corpus con mayor variedad de registros (Davis 2006).

3.2 Proceso de anotación

Las frases anotadas pasan por un proceso de revisión para homogeneizar las anotaciones (ver apartado 3.3). Hasta este momento se ha estudiado la tipología de errores, se ha establecido una metodología semiautomática de revisión y se ha revisado un 10% del corpus. El proyecto finaliza a finales del 2006. Para entonces se prevé haber revisado el 60% del corpus manualmente, y el 100% automáticamente.

En una primera etapa del proceso de anotación, cuando los criterios para llevar a cabo la anotación no estaban consolidados, se repartieron entre los cuatro anotadores iniciales una serie de ejemplos comunes, que anotaría cada uno por separado. Estas anotaciones de frases comunes se compararon con el fin de descubrir cuáles eran los puntos en que la subjetividad y la falta de un criterio uniforme creaban mayores divergencias o confusiones. El estudio de estas comparaciones ha permitido establecer una documentación que prevé y soluciona los problemas más frecuentes que pueden surgir en el proceso de anotación, lo cual asegura una mayor consistencia en las anotaciones, minimizando la influencia de los criterios subjetivos de cada anotador. El mayor desacuerdo se detectó inicialmente en la distinción entre argumentos y adjuntos, en el uso del rol *iniciador* y en la distinción entre los diferentes tipos de objetos preposicionales previstos (Vázquez et al. 2005).

3.3 Proceso de revisión

Como en todos los corpus anotados manualmente, la anotación humana no se libra de un cierto porcentaje de error. En el caso del corpus SenSem, la naturaleza de los errores es diversa. Afecta en mayor grado a los niveles de descripción con una granularidad de categorías más fina: las categorías morfosintácticas y las funciones sintácticas. También encontramos errores, aunque en menor medida, en la segmentación de oraciones a anotar, la interpretación oracional y la estructura eventiva. Cuantitativamente, el 17% de las oraciones tienen por lo menos un error, y la cantidad de errores es un 25% sobre el número total de oraciones. Para una descripción más detallada de los errores encontrados en corpus, véase Alonso et al. (2006).

Para corregir estos errores se han establecido procesos de revisión automática y manual. En este último intervienen dos jueces y, por último, aún intercede un revisor final.

Para sistematizar el proceso de revisión, se han detectado diversos errores que hemos atribuido a diferentes causas como lapsus del anotador, categorías con definición poco específica en los criterios o inherentemente infraespecificadas o bien errores de concepción gramatical.

El proceso de revisión se organiza en dos fases. Primero, se aplican rutinas de detección y corrección automática de errores. Estas rutinas detectan inconsistencias diversas entre los diferentes niveles de anotación, y, en general, inconsistencias detectables por correspondencia de patrones simple, por ejemplo:

- si una oración antiagentiva tiene un objeto directo
- si un sintagma adverbial está etiquetado como objeto preposicional
- si un sintagma preposicional no empieza por preposición
- si no se ha marcado el núcleo de un argumento, etc.

Mediante estos procedimientos de correspondencia de patrones podemos solucionar la mayoría de errores de lapsus de anotador o problemas de comprensión de los conceptos gramaticales.

En segundo lugar se revisa el resultado de la corrección automática mediante un procedimiento manual, en el que se corrigen también incoherencias interpretativas, no detectables automáticamente, como por ejemplo asignaciones incorrectas de rol semántico, de sentido verbal o de interpretación oracional.

4 Herramientas

A lo largo de todo el proyecto ha sido necesaria la creación de diferentes herramientas tanto para la anotación como para la difusión de los resultados y la confección de una BD léxica. Dichas herramientas se encuentran a disposición de la comunidad científica.

4.1 Interfaz de anotación

La anotación de las oraciones se ha realizado de forma semiautomática mediante una interfaz gráfica en la que se distinguen claramente los tres niveles: semántica oracional, léxico verbal y de constituyentes. El sistema de anotación funcionó primero en local y actualmente está implementado como servicio web.

En la figura 1 se muestra la pantalla que se le presenta al anotador para asignar el sentido verbal que más se adecua al ejemplo.

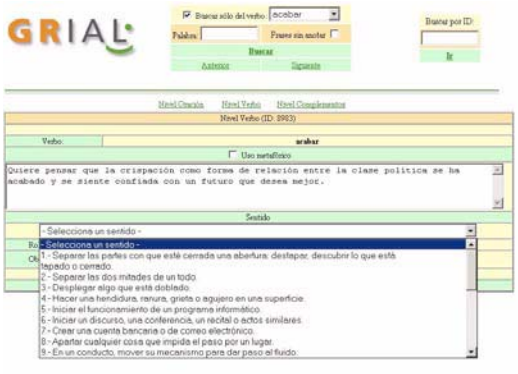

Figura 1: Anotación del sentido verbal

El inventario de sentidos, como se ha mencionado, es anterior al proceso de anotación, pero es modificado siempre que los ejemplos encontrados en el corpus demuestren la existencia de un sentido no contemplado, o bien, cuando los ejemplos son difíciles de clasificar porque se da confusión entre sentidos, lo que provoca que algunos de ellos acaben fusionándose. Además, si en alguna frase el sentido verbal se utiliza metafóricamente y este uso no se considera suficientemente habitual como para crear un nuevo sentido, se marca como uso metafórico del sentido inicial.

Tras la selección del sentido, la interfaz muestra el patrón prototípico de roles semánticos que *a priori* participan en los ejemplos correspondientes al sentido seleccionado. El anotador toma esta información y, delimitando los constituyentes, les asigna los roles semánticos, excepto a aquellos que no son considerados argumentales. Como esa lista de roles también ha sido asignada al sentido *a priori*, también se modifica si los ejemplos del corpus proporcionan una evidencia no contemplada inicialmente. Todas las modificaciones de la información preestablecida han sido consensuadas.

La anotación manual de los roles semánticos permite anotar automáticamente otro tipo de información asociada a cada constituyente: la categoría y la función que con más frecuencia están vinculadas a ese papel son preseleccionadas por el sistema y el anotador debe sólo validar o modificar esta información.

También se marcan los núcleos de los constituyentes, y su posible uso metafórico. Finalmente, el anotador determina la información sobre la semántica oracional del ejemplo, ya sea anticausativa, reflexiva, etc., e indica los comentarios que considere oportunos para procesos futuros como la revisión y la corrección.

4.2 Interfaz de consulta

El corpus anotado se puede consultar vía web (http://grial.uab.es/search), mediante una interfaz que permite acceder a la información anotada del corpus (para una descripción detallada de la interfaz, véase Fernández et al. 2006).

Como se puede observar en la figura 2, la herramienta de explotación creada permite consultar información específicamente para un sentido verbal y combinar consultas de diferentes niveles (léxico, de constituyentes y oracional).

En el nivel de constituyentes, se puede requerir simultáneamente información de hasta un máximo de tres sintagmas de la misma oración.

Además, la forma de visualizar los resultados de una consulta pretende facilitar el acceso rápido a la información, aún si se ha requerido más de un objeto de búsqueda.

A partir de los resultados de la consulta realizada (figura 3), también se pueden examinar otros aspectos de la anotación que no se han solicitado para obtener la anotación completa de la oración. A esta utilidad se accede a través del botón “anotación”.

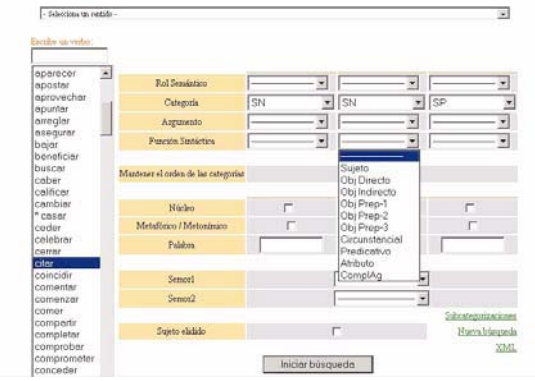

Figura 2: Interfaz de búsqueda

Esta interfaz de consulta incluye otra específica para la consulta de los patrones de subcategorización asociados a un sentido. Las distintas posibilidades de la realización de cada esquema y de la combinación de diferentes informaciones anotadas han dado lugar a una lista muy amplia de dichos patrones. Al realizar la consulta se obtiene también la información sobre la

frecuencia de cada esquema y sobre las frases asociadas a cada uno de ellos.

Figura 3: Resultados de la búsqueda

4.3 BD léxica: versión preliminar

En esta interfaz se podrá consultar una versión preliminar de lo que será la BD léxica, la cual contendrá tanto información relativa al verbo (sentido) como a las construcciones en las que participa. Falta definir todavía el formato definitivo de dicha BD, pero en principio se prevén incluir los datos que se exponen a continuación.

Respecto al verbo, incluimos:

- la definición
- el correspondiente synset de EuroWordNet
- la estructura eventiva
- el conjunto de papeles semánticos asociado.

Durante la fase de adquisición, se pretende extraer también información relativa a las restricciones de selección (a partir de la anotación de los núcleos y su confrontación con una jerarquía) y a las preposiciones.

En cuanto a las construcciones en que participa (a partir de los datos obtenidos del corpus), la información que se incluye es la siguiente:

- patrones de subcategorización simples (categorías sintagmáticas generales) y ampliados (categorías sintagmáticas específicas, funciones sintácticas y roles semánticos)
- semántica de cada construcción con respecto a la focalización de los constituyentes (anticausatividad, antiagentividad, etc.), la reflexividad o reciprocidad u otras como las construcciones de dativo de interés
- ejemplos anotados del corpus

Una de las fases prevista que estamos llevando a cabo es la conexión de esta BD con la totalidad del corpus anotado. La unión de los dos recursos constituirá lo que hemos denominado un "banco de datos del español". Se puede consultar una versión preliminar del mismo en: http://grial.uab.es/adquisicio.

5 *Conclusiones y líneas futuras*

El banco de datos que hemos presentado constituye una importante fuente de información lingüística útil para diversas aplicaciones de procesamiento de lenguaje natural, así como para la investigación lingüística. Creemos que el hecho de que el corpus se esté anotando en diversos niveles de representación lingüística incrementa su valor y su versatilidad. Esto puede ser especialmente útil para aplicaciones en el campo de la comprensión automática y en el de la representación semántica, así como para sistemas que utilicen técnicas de aprendizaje automático.

Actualmente estamos en el último año de desarrollo del proyecto, dedicado a la finalización del proceso de anotación, la corrección de la misma y la extracción y adquisición de los datos anotados para la construcción de la base de datos léxica SenSem.

Además del recurso en sí, a partir del proceso de anotación se han establecido los criterios de anotación, se han analizado los errores y establecido una tipología y una serie de procedimientos automáticos para su corrección. Consideramos que estos estudios son útiles para otro tipo de aplicaciones o estudios en el ámbito del Procesamiento de lenguaje Natural.

Referencias

Alonso, L., I. Castellón, N. Tincheva. 2006. Detección automática de errores en el Corpus Sensem. Congreso de la Asociación Española de Lingüística Aplicada.

Davies, M. 2006. A Frequency dictionary of Spanish: core vocabulary for learners. New York-London . Routledge.

Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara, M. Kamel. 2002. The VOLEM Project:

a Framework for the Construction of Advanced Multilingual Lexicons. Proceedings of the Language Engineering Conference, p. 89-98. ISBN: 0-7695-1885-0/02.

Fernández, A., G. Vázquez, D. Teruel. 2006. Interfaz de explotación del corpus SenSem. Congreso de la Asociación Española de Lingüística Aplicada.

Garcia de Miguel, J.M. and S. Comesaña. 2004. Verbs of Cognition in Spanish: Constructional Schemas and Reference Points, in A. Silva, A. Torres, M. Gonçalves (eds) Linguagem, Cultura e Cogniçao: Estudos de Linguística Cognitiva, Almedina, pp. 399-420.

Palomar, M., M. Civit, A. Díaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A., Martí, B. Navarro. 2004. 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. Revista de la SEPLN, Barcelona. pp. 81-87.

Subirats-Rüggeberg, C. y M. R. L. Petruck. 2003. Surprise: Spanish FrameNet! Presentation at the Workshop on Frame Semantics, Proceedings of the International Congress of Linguists, Praga.

Vossen, Piek. 1999. EuroWordNet General Document. Version 3, Final. University of Amsterdam.

Vázquez, G., A. Fernández, L. Alonso. 2005. Description of the Guidelines for the Syntactico-semantic Annotations of a Corpus in Spanish. Angelova, G., K. Bontcheva, R. Mitkov, N. Nicolov (ed.), International Conference Recent Advances in Natural Language. Shoumen (Bulgaria), p. 603-607.